STAT 113
Simple Linear Regression I

Colin Reimer Dawson

Oberlin College

September 20, 2017

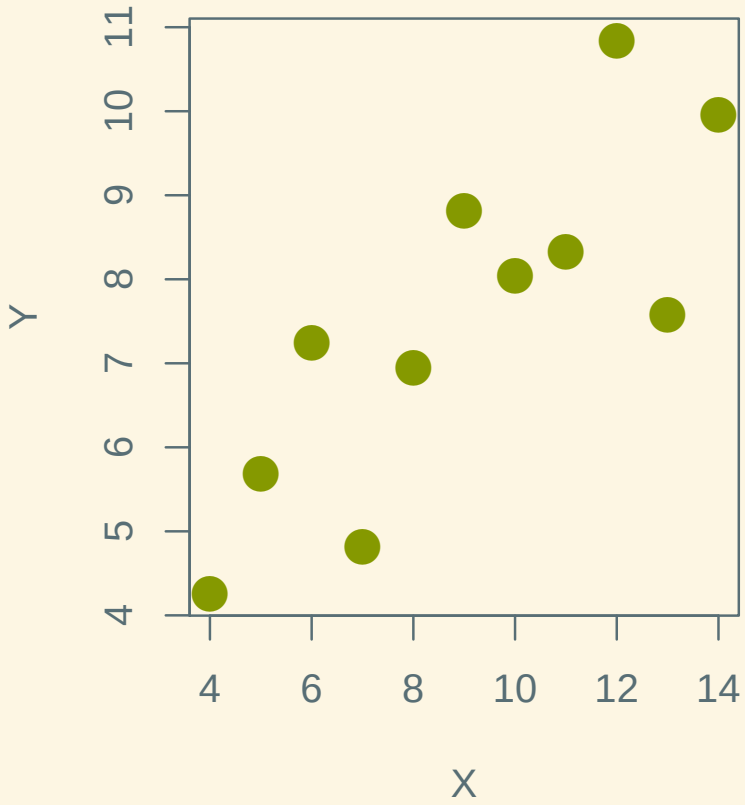
Outline

Prediction
- Linear Models
- Finding the “Best” Line”

Prediction

- Correlations give us a **description** of the relationship between two numeric variables.
- However, when two variables are related, we can go further and use knowledge of one to make **predictions** about the other.
- Examples:
  - Use height to predict blood pressure
  - Use economic indicators to predict stock prices
  - Use credit score to predict probability of default on a loan
  - Use biomarkers to predict disease progression

What's a Good Prediction?

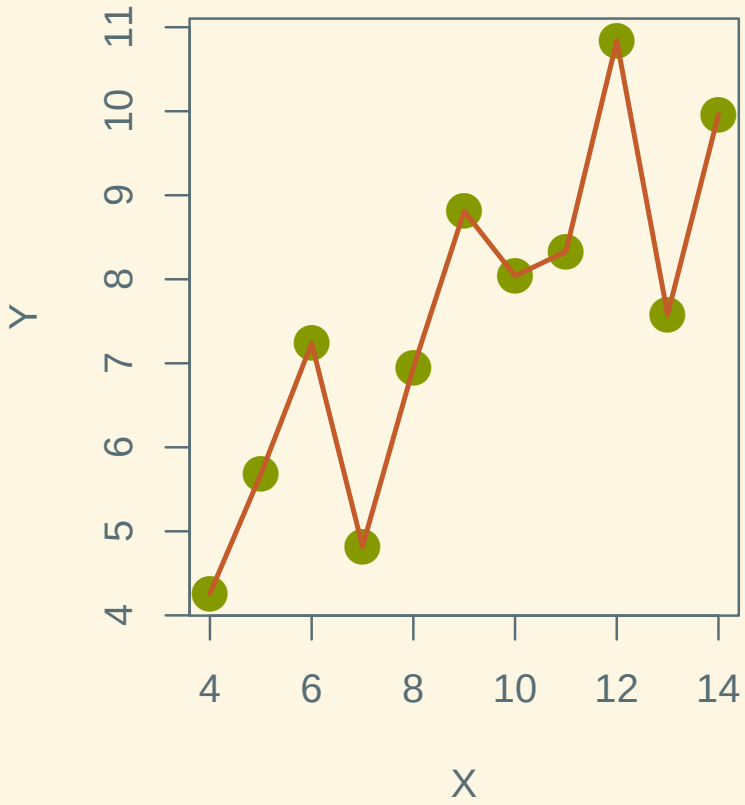


- Suppose I have this data.
- What would be a good prediction if I get a new X value of 12?
- What about an X value of 5.5?

Modeling relationships with a function

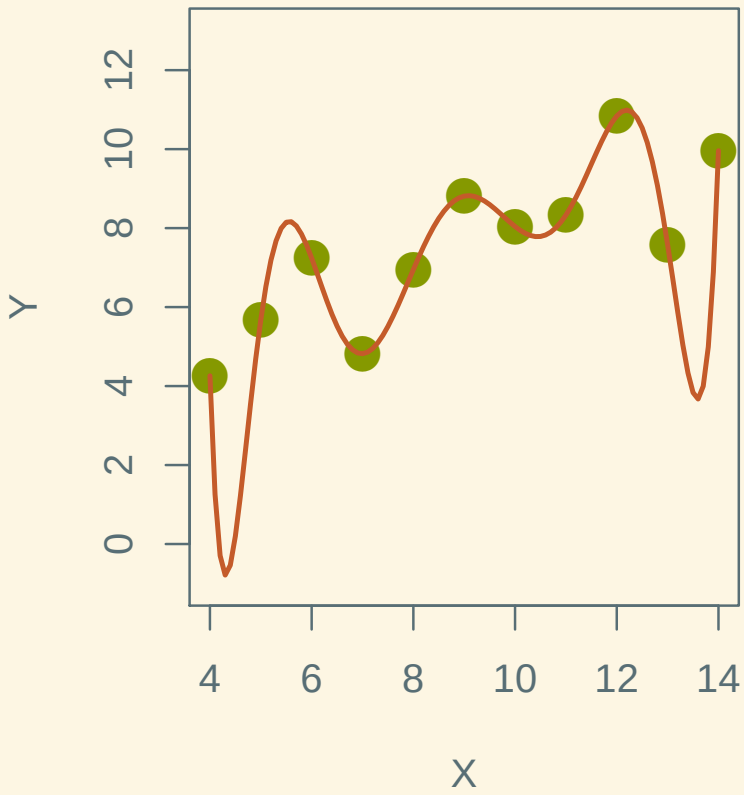
- We can capture all of our predictions by writing the y variable as a **function** of the x variable
- Examples:
  - $f(x) = x^2$
  - $f(x) = 1.6x + 20$
  - $f(x) = 5\cos(2\pi x)$

What's a Good Prediction?



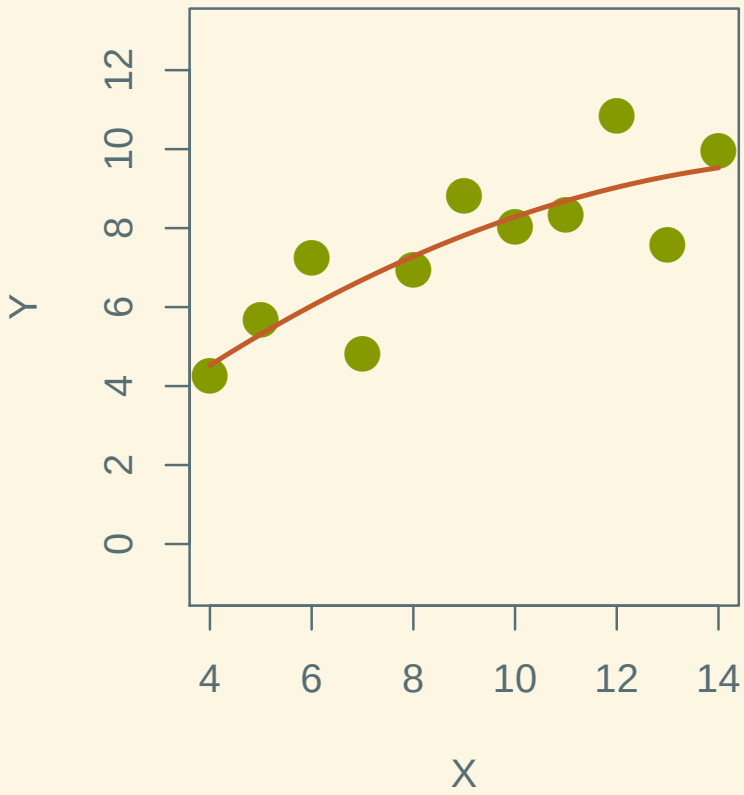
How about this function?

What's a Good Prediction?



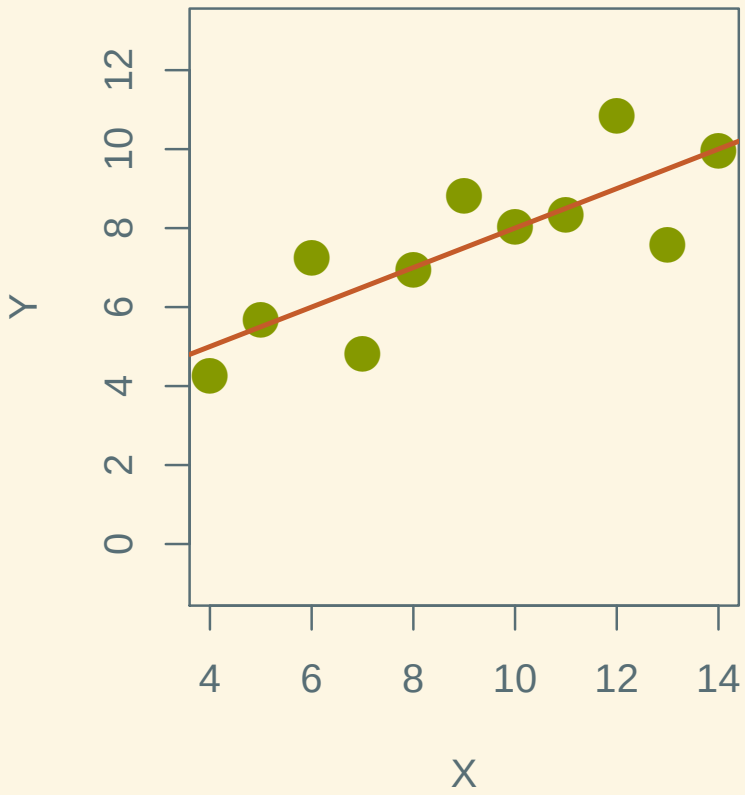
Or this?

What's a Good Prediction?



- What about this?
- There's a tradeoff between how well we can fit the data and how simple our **model** (i.e., prediction function) is.

What's a Good Prediction?



- Pretty much the simplest model we can have is a straight line.
- Two things determine what line we have:
  - The intercept
  - The slope

Intercept Slope Form

- The general equation for a line is:

$$f(x) = a + bx$$

- The intercept and slope are the **parameters** of our regression model.
- In statistics notation, we write $\hat{y}$ ("y hat") to represent a *predicted* (or fitted) value.
- Given a value x_i, we predict using:

$$\hat{y} = \hat{a} + \hat{b}x_i$$

the equation of the "best" line through our sample data

Hat Notation

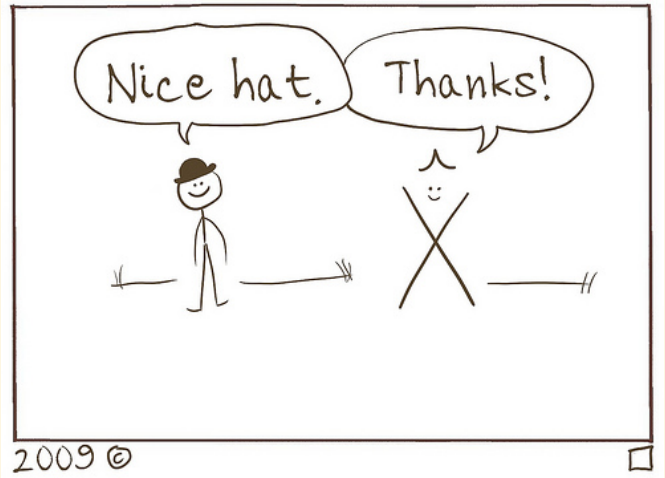


Figure: Source: `brownsharpie.com`

Example: 2016 Presidential Election

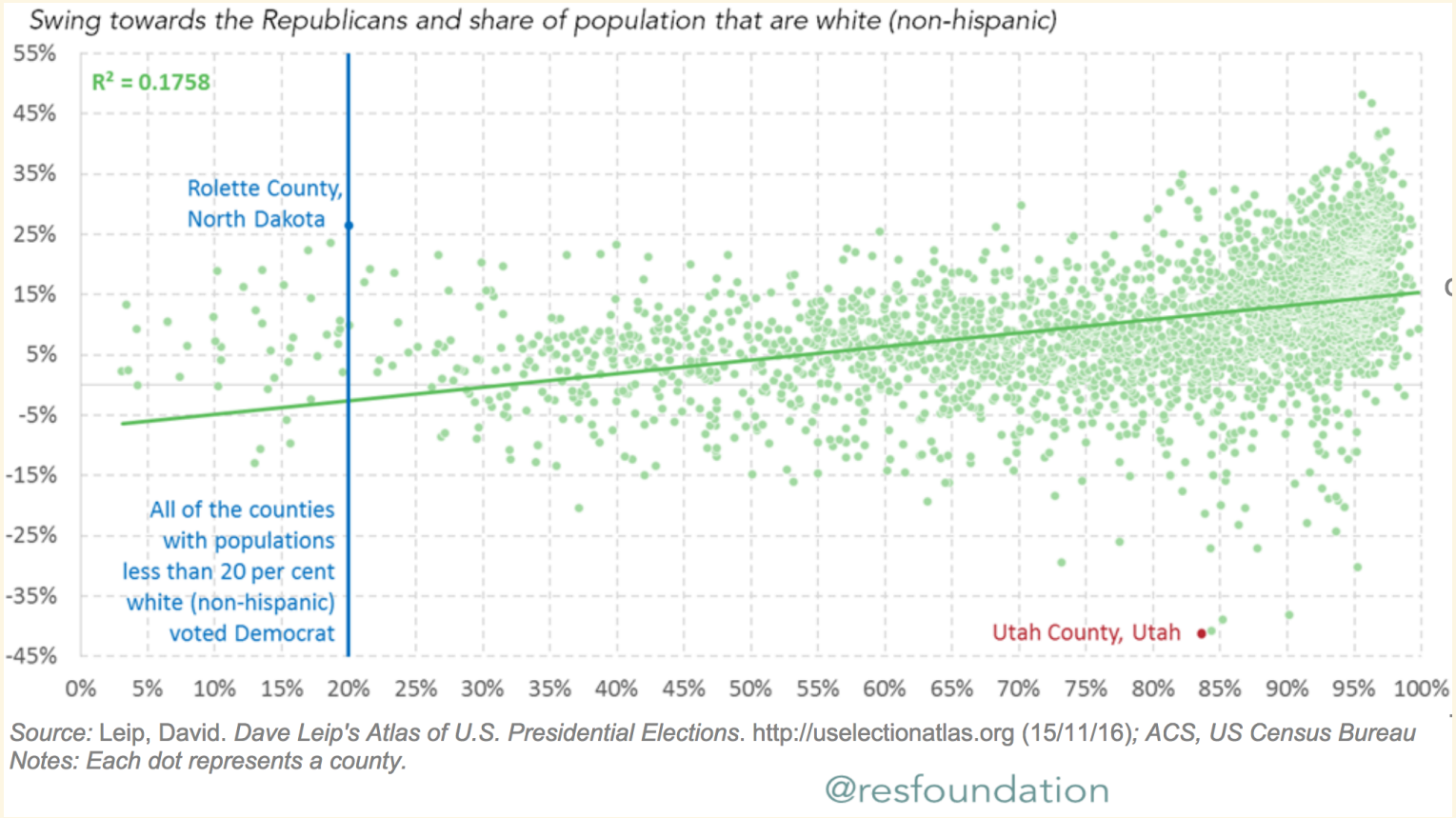


Figure: Difference in share of county voting Republican, 2016 minus 2012, as a function of percent of the county that is white (Source: Resolution Foundation)

Example: 2016 Presidential Election

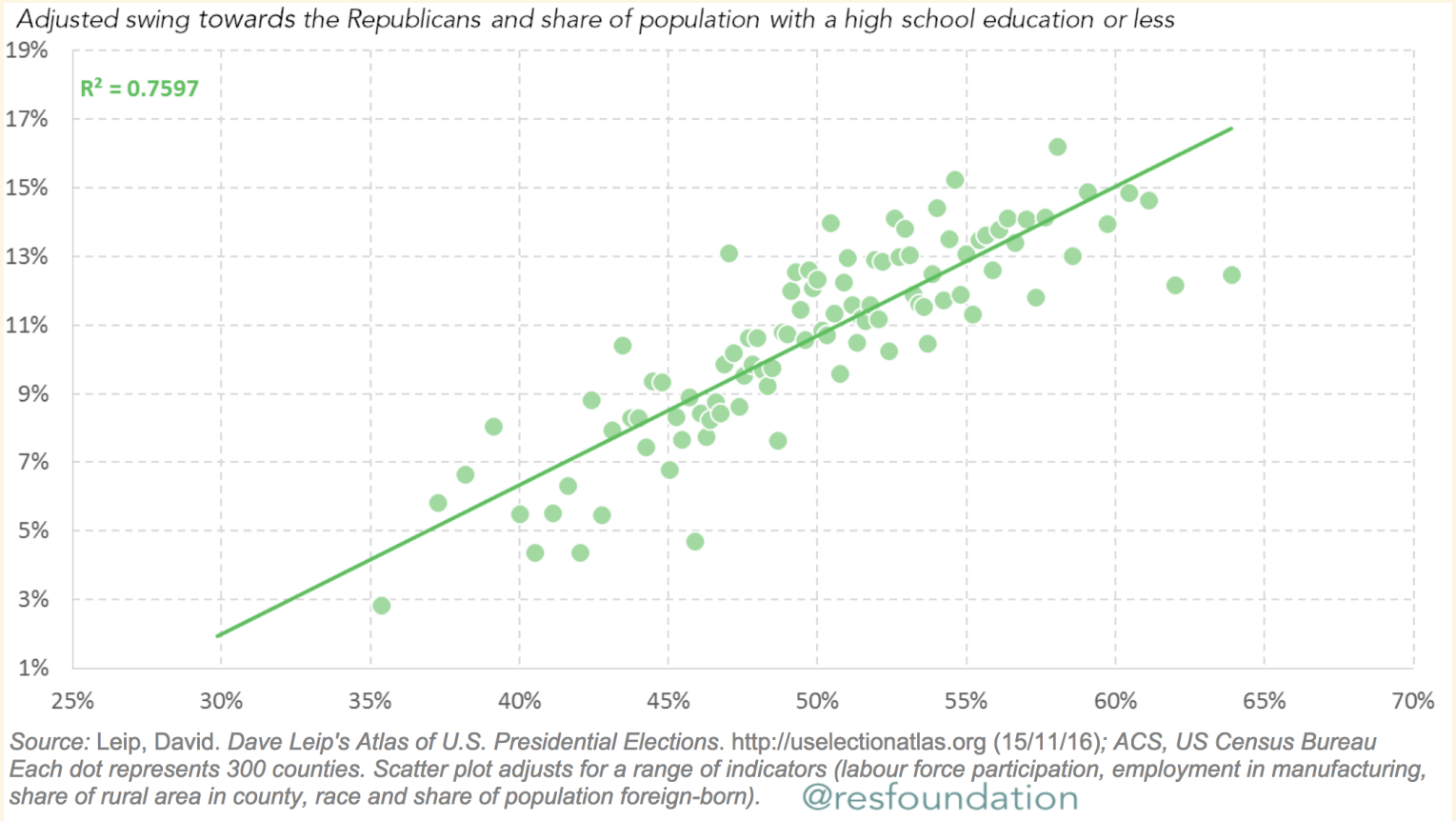


Figure: Difference in share of county voting Republican, 2016 minus 2012, as a function of percent of the county that has HS education or less. Each dot is 300 counties. (Source: Resolution Foundation)

Example: 2016 Presidential Election

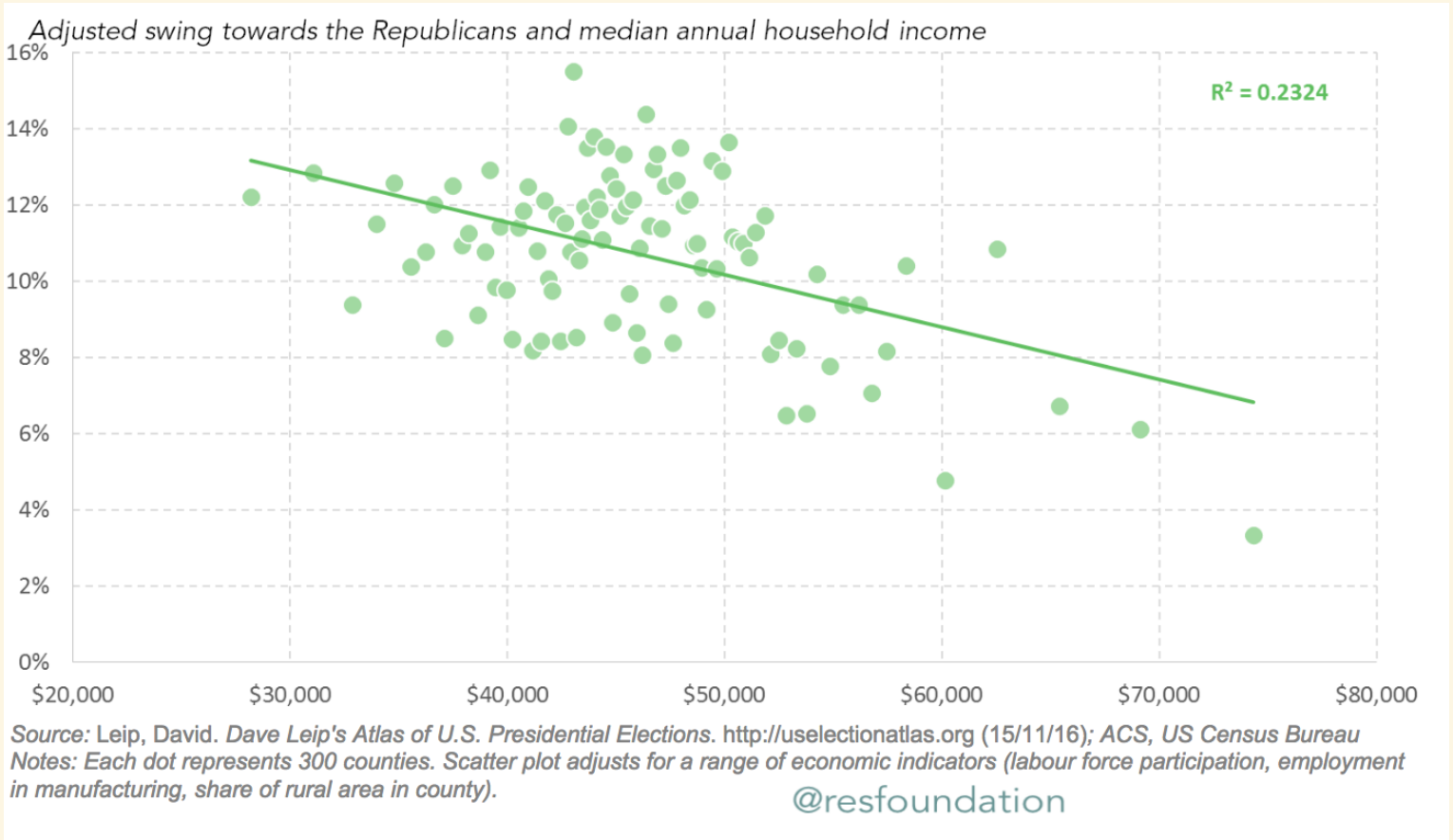


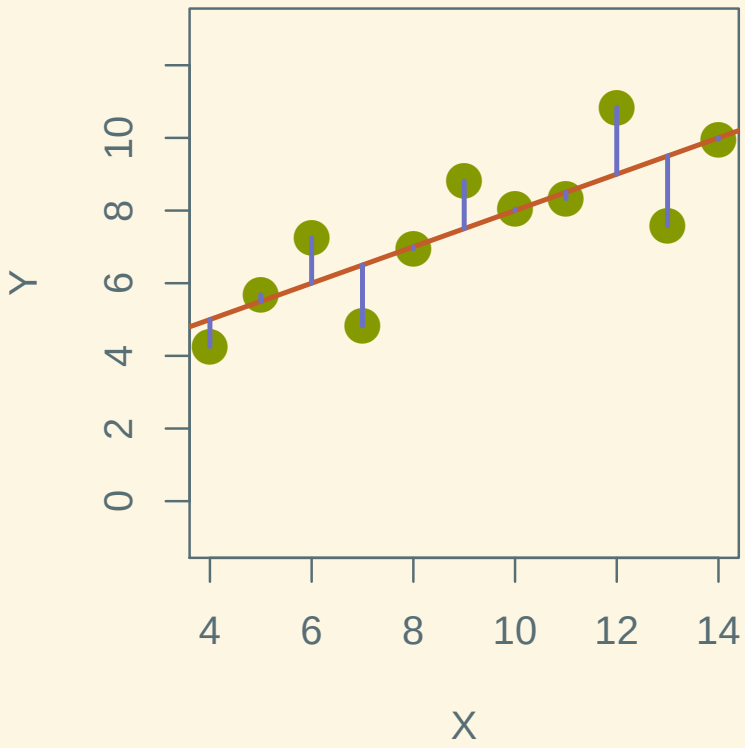
Figure: Difference in share of county voting Republican, 2016 minus 2012, as a function of median county income Each dot is 300 counties. (Source: Resolution Foundation)

Systematic vs. Random

- We can split up each y value into two parts: a systematic (predictable) part and a "random" part.
- That is, we can write, for the y coordinate of the ith data point:

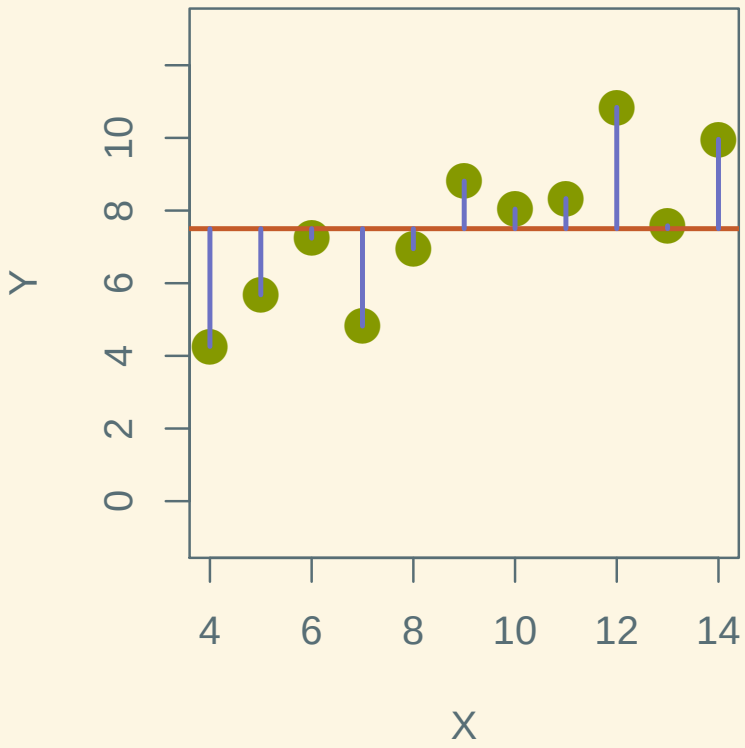
$$y_i = \hat{y}_i + \text{Error}_i$$

What's a Good Prediction?



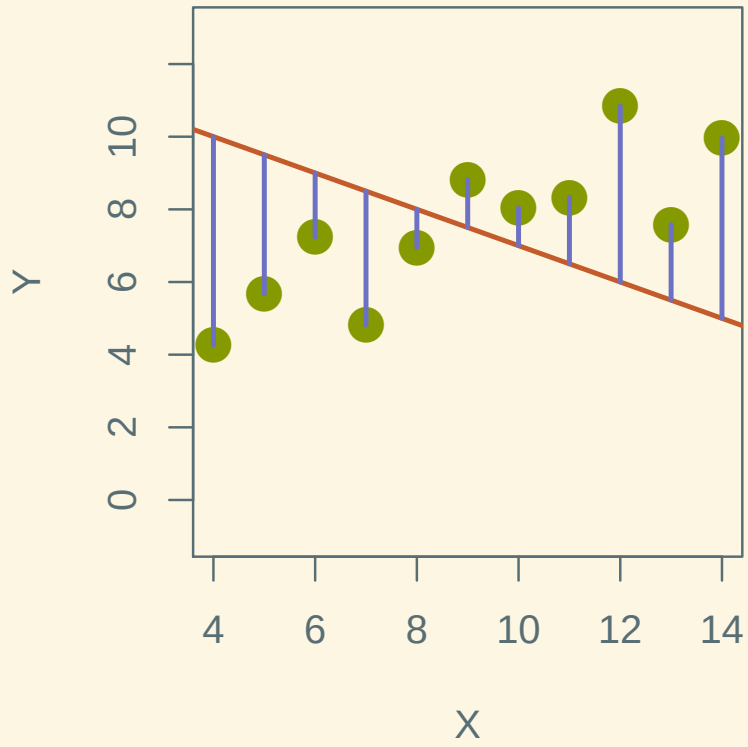
Every line will have a different set of errors associated with it.

What's a Good Prediction?



Every line will have a different set of errors associated with it.

What's a Good Prediction?



- Every line will have a different set of errors associated with it.
- Which is best?
- Intuitively, we want to minimize the overall "distance" between the line and the points.

The Prediction Equation

Prediction Function

$$\hat{y}_i = \hat{a} + \hat{b}_1 x_i$$

Pick $\hat{a}$ and $\hat{b}$ that minimize the total distance. This is a calculus problem that the computer solves for us.

Choose $\hat{a}$ and $\hat{b}$ so as to minimize "distance between sample and line":

$$SSE = \sqrt{\sum_i (y_i - \hat{y}_i)^2} = \sqrt{\sum_i (y_i - \hat{a} - \hat{b}x_i)^2}$$

Handout: Correlation and Regression

Handout: Outliers, Correlation and Regression